Topic 6: - Topic Analysis

Clarissa Boyajian

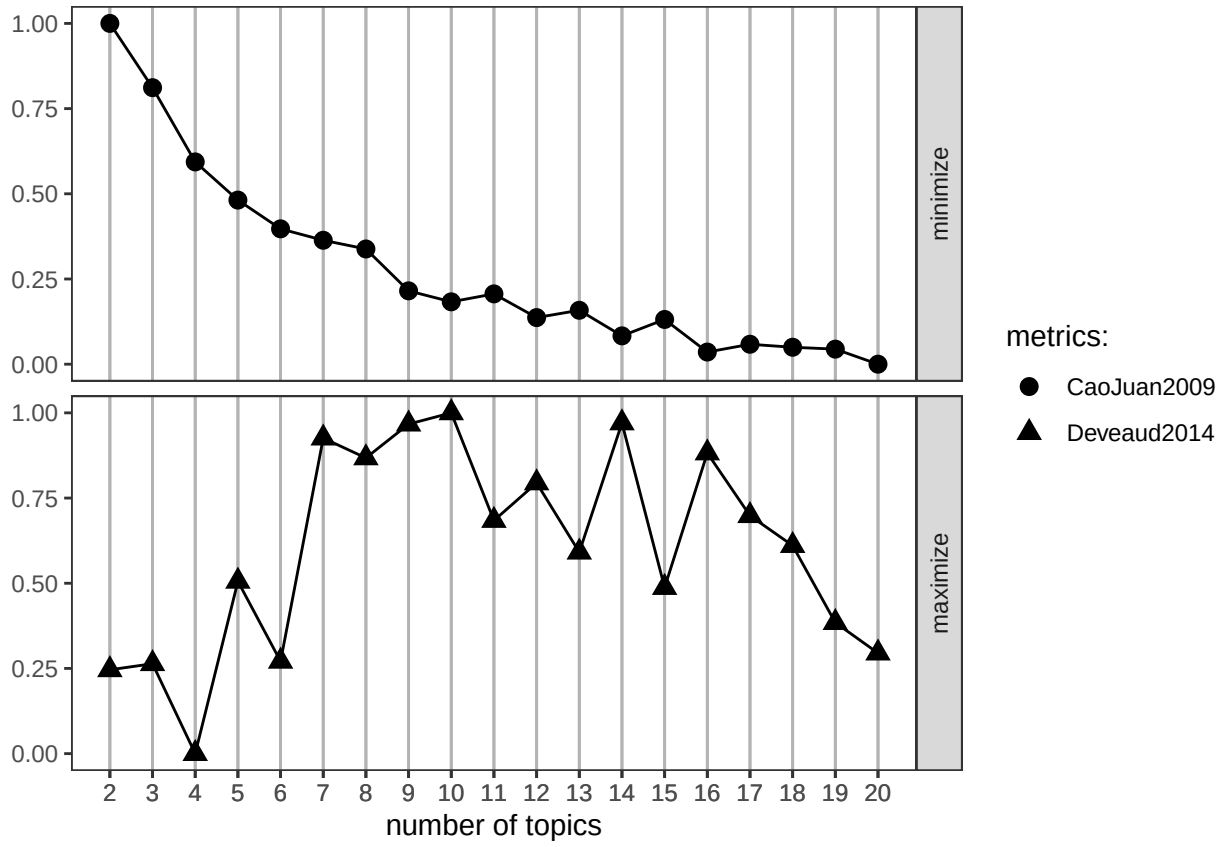
2022-05-10

```
## -- read in, clean, and wrangle data -- ##
comments_df<-read_csv(here("data/comments_df.csv"))

epa_corp <- corpus(x = comments_df, text_field = "text")
epa_corp.stats <- summary(epa_corp)
tokens <- tokens(epa_corp, remove_punct = TRUE, remove_numbers = TRUE)

# project-specific stop words
add_stops <- c(stopwords("en"),"environmental", "justice", "ej", "epa", "public", "comment")
tokens_1 <- tokens_select(tokens, pattern = add_stops, selection = "remove")

doc_feature_matric_common <- dfm(tokens_1, tolower = TRUE)
doc_feature_matric <- dfm_wordstem(doc_feature_matric_common)
doc_feature_matric <- dfm_trim(doc_feature_matric, min_docfreq = 2)

# remove rows (docs) with all zeros
select_idx <- slam::row_sums(doc_feature_matric) > 0
doc_feature_matric <- doc_feature_matric[select_idx, ]
```

Calculate value of k that is most likely

```
# calculate what initial value of k is most likely
result <- FindTopicsNumber(
  doc_feature_matric,
  topics = seq(from = 2, to = 20, by = 1),
  metrics = c("CaoJuan2009",  "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  verbose = TRUE
)
```

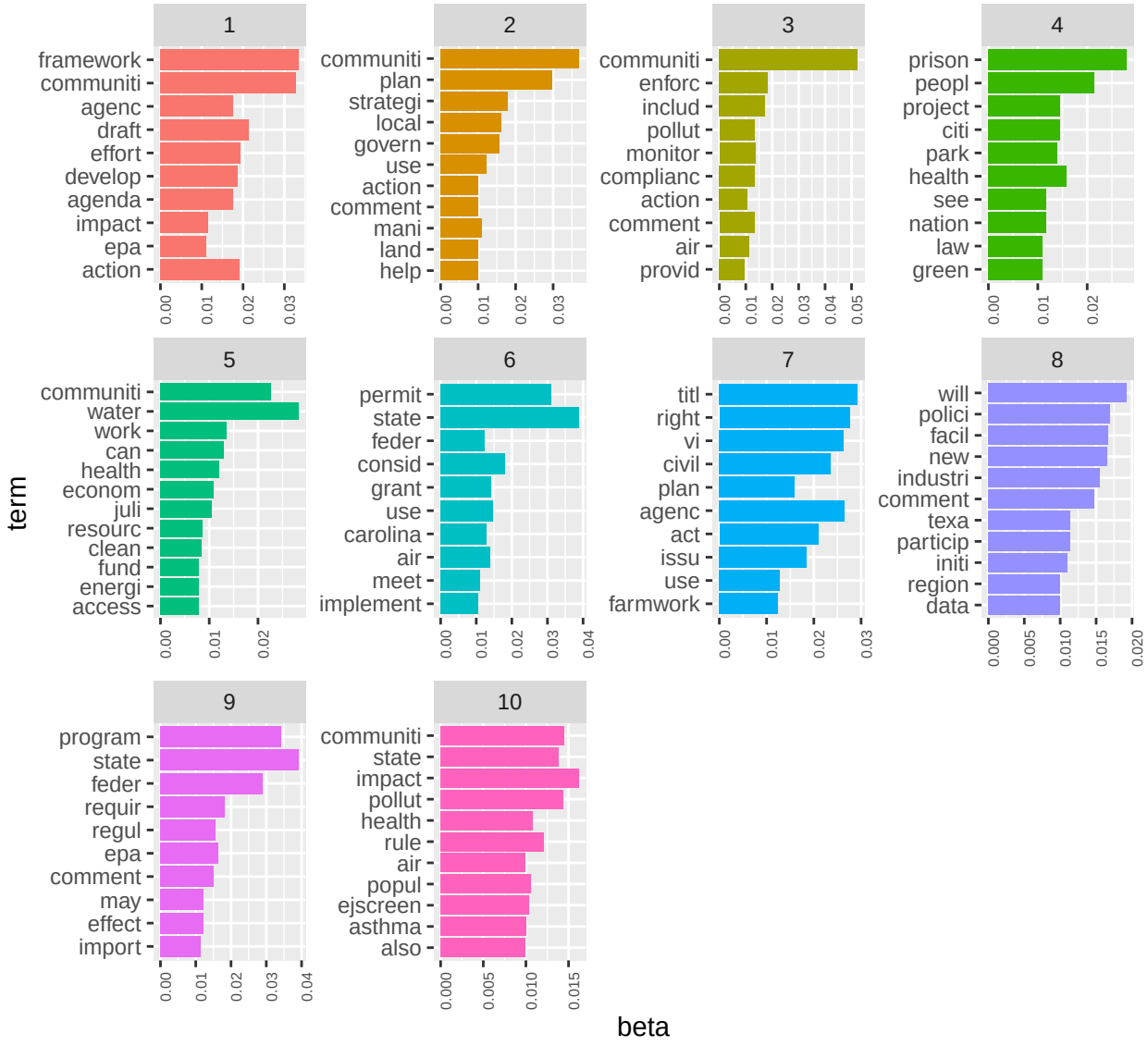
```
FindTopicsNumber_plot(result)
```



Model 1 (k = 10)

```
k <- 10

topicModel_k10 <- LDA(doc_feature_matric, k,
                      method = "Gibbs",
                      control = list(iter = 500, verbose = 25))

tmResult_k10 <- posterior(topicModel_k10)
theta_k10 <- tmResult_k10$topics
beta_k10 <- tmResult_k10$terms # probability of each term in each topic
vocab_k10 <- (colnames(beta_k10))

comment_topic_k10 <- tidy(topicModel_k10, matrix = "beta")

top_terms_k10 <- comment_topic_k10 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

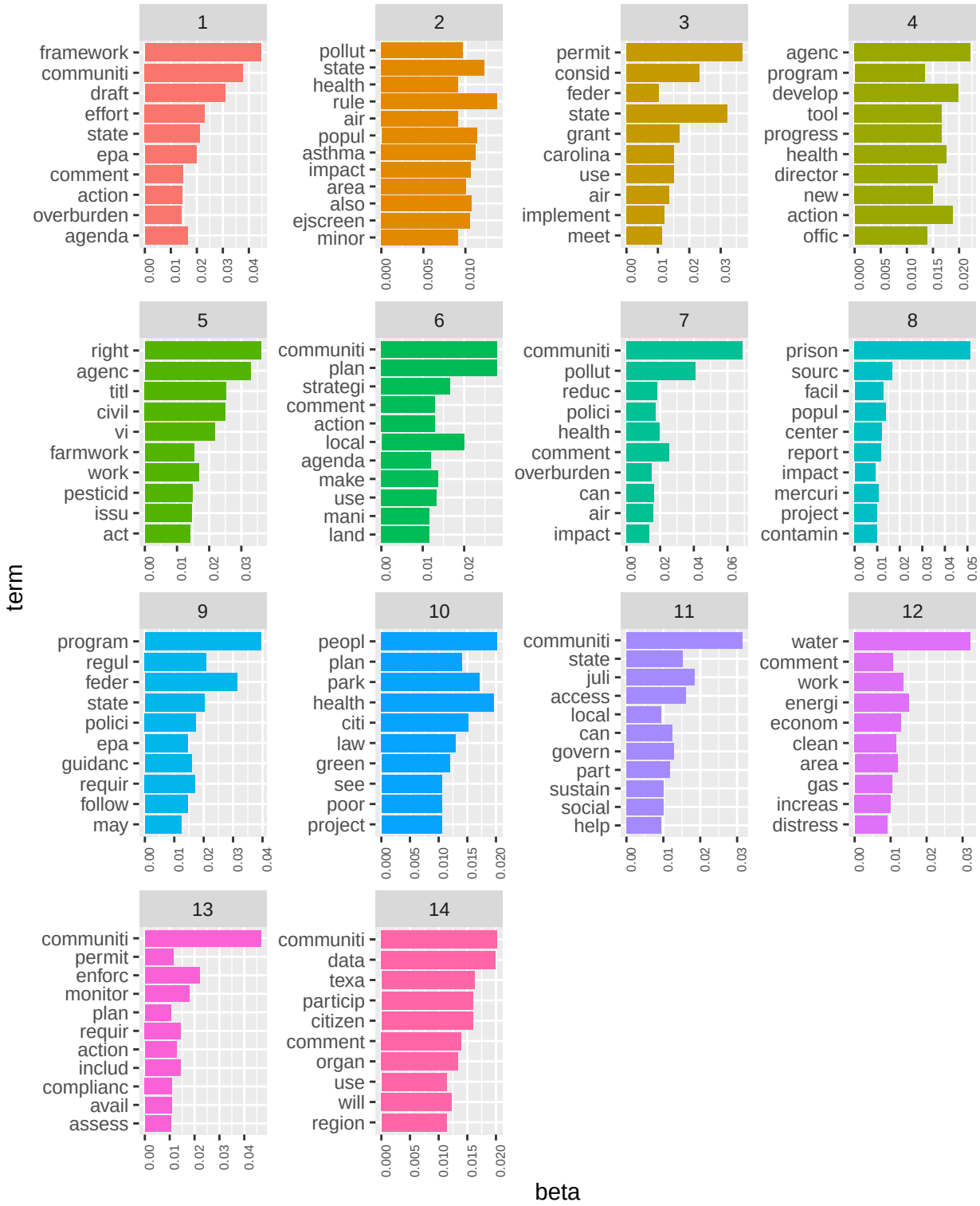
```
top_terms_k10 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 6))
```



Model 2 (k = 14)

```
k <- 14

topicModel_k14 <- LDA(doc_feature_matric, k,
                      method = "Gibbs",
                      control = list(iter = 500, verbose = 25))

tmResult_k14 <- posterior(topicModel_k14)
theta_k14 <- tmResult_k14$topics
beta_k14 <- tmResult_k14$terms # probability of each term in each topic
vocab_k14 <- (colnames(beta_k14))

comment_topic_k14 <- tidy(topicModel_k14, matrix = "beta")

top_terms_k14 <- comment_topic_k14 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

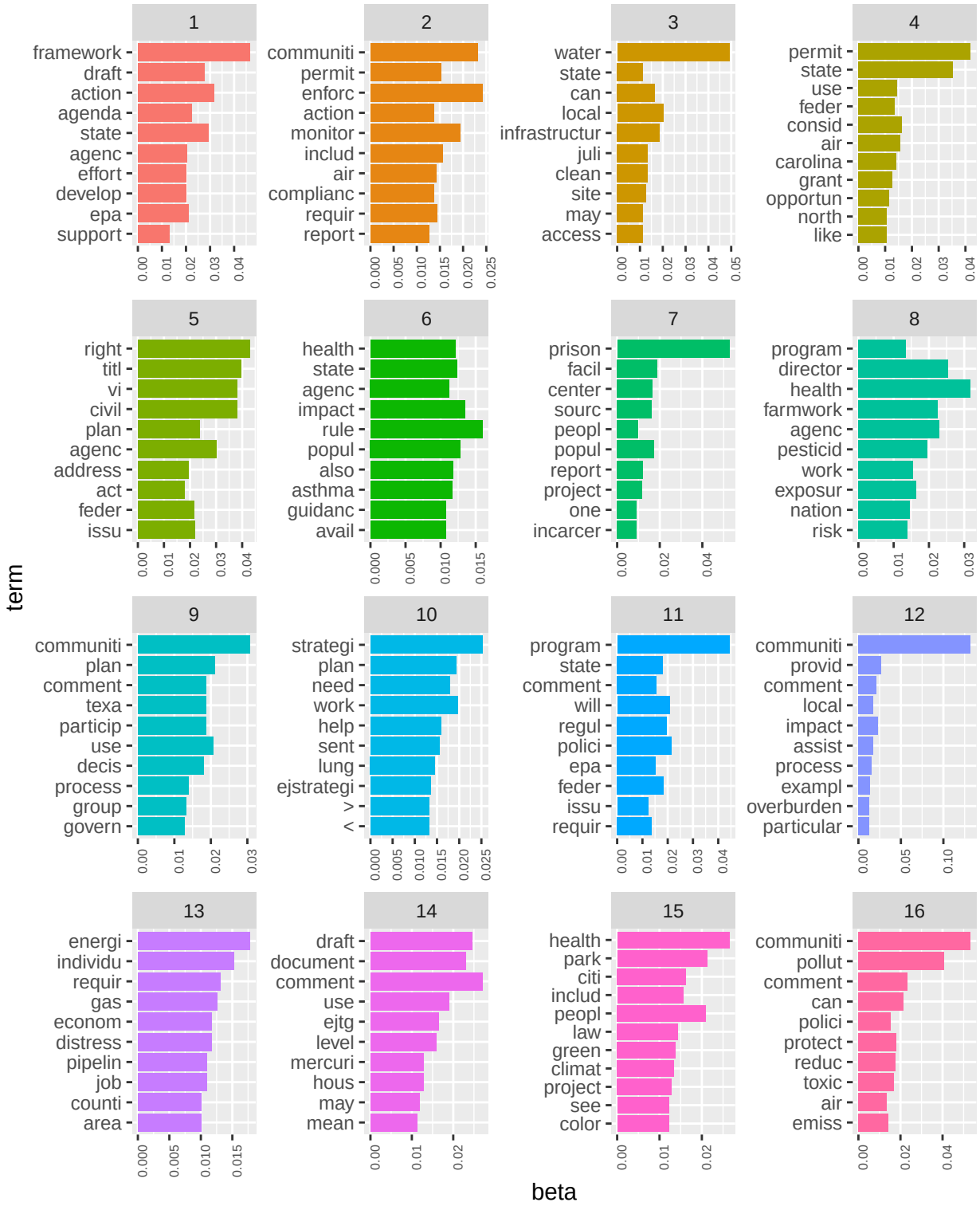
```
top_terms_k14 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 6))
```


1
framework
communiti
draft
effort
state
epa
comment
action
overburden
agenda
0.00
0.01
0.02
0.03
0.04
2
pollut
state
health
rule
air
popul
asthma
impact
area
also
ejscreen
minor
0.000
0.005
0.010
3
permit
consid
feder
state
grant
carolina
use
air
implement
meet
0.00
0.01
0.02
0.03
4
agenc
program
develop
tool
progress
health
director
new
action
offic
0.000
0.005
0.010
0.015
0.020
5
right
agenc
titl
civil
vi
farmwork
work
pesticid
issu
act
0.00
0.01
0.02
0.03
6
communiti
plan
strategi
comment
action
local
agenda
make
use
mani
land
0.00
0.01
0.02
7
communiti
pollut
reduc
polici
health
comment
overburden
can
air
impact
0.00
0.02
0.04
0.06
8
prison
sourc
facil
popul
center
report
impact
mercuri
project
contamin
0.00
0.01
0.02
0.03
0.04
0.05
9
program
regul
feder
state
polici
epa
guidanc
requir
follow
may
0.00
0.01
0.02
0.03
0.04
10
peopl
plan
park
health
citi
law
green
see
poor
project
0.000
0.005
0.010
0.015
0.020
11
communiti
state
juli
access
local
can
govern
part
sustain
social
help
0.00
0.01
0.02
0.03
12
water
comment
work
energi
econom
clean
area
gas
increas
distress
0.00
0.01
0.02
0.03
13
communiti
permit
enforc
monitor
plan
requir
action
includ
complianc
avail
assess
0.00
0.01
0.02
0.03
0.04
14
communiti
data
texa
particip
citizen
comment
organ
use
will
region
0.000
0.005
0.010
0.015
0.020
term
beta

Model 3 (k = 16)

```
k <- 16

topicModel_k16 <- LDA(doc_feature_matric, k,
                      method = "Gibbs",
                      control = list(iter = 500, verbose = 25))

tmResult_k16 <- posterior(topicModel_k16)
theta_k16 <- tmResult_k16$topics
beta_k16 <- tmResult_k16$terms # probability of each term in each topic
vocab_k16 <- (colnames(beta_k16))

comment_topic_k16 <- tidy(topicModel_k16, matrix = "beta")

top_terms_k16 <- comment_topic_k16 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

```
top_terms_k16 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 6))
```


1
framework
draft
action
agenda
state
agenc
effort
develop
epa
support
2
communiti
permit
enforc
action
monitor
includ
air
complianc
requir
report
3
water
state
can
local
infrastructur
juli
clean
site
may
access
4
permit
state
use
feder
consid
air
carolina
grant
opportun
north
like
5
right
titl
vi
civil
plan
agenc
address
act
feder
issu
6
health
state
agenc
impact
rule
popul
also
asthma
guidanc
avail
7
prison
facil
center
sourc
peopl
popul
report
project
one
incarcer
8
program
director
health
farmwork
agenc
pesticid
work
exposur
nation
risk
9
communiti
plan
comment
texa
particip
use
decis
process
group
govern
10
strategi
plan
need
work
help
sent
lung
ejstrategi
v
<
11
program
state
comment
will
regul
polici
epa
feder
issu
requir
12
communiti
provid
comment
local
impact
assist
process
exampl
overburden
particular
13
energi
individu
requir
gas
econom
distress
pipelin
job
counti
area
14
draft
document
comment
use
ejtg
level
mercuri
hous
may
mean
15
health
park
citi
includ
peopl
law
green
climat
project
see
color
16
communiti
pollut
comment
can
polici
protect
reduc
toxic
air
emiss
term
beta

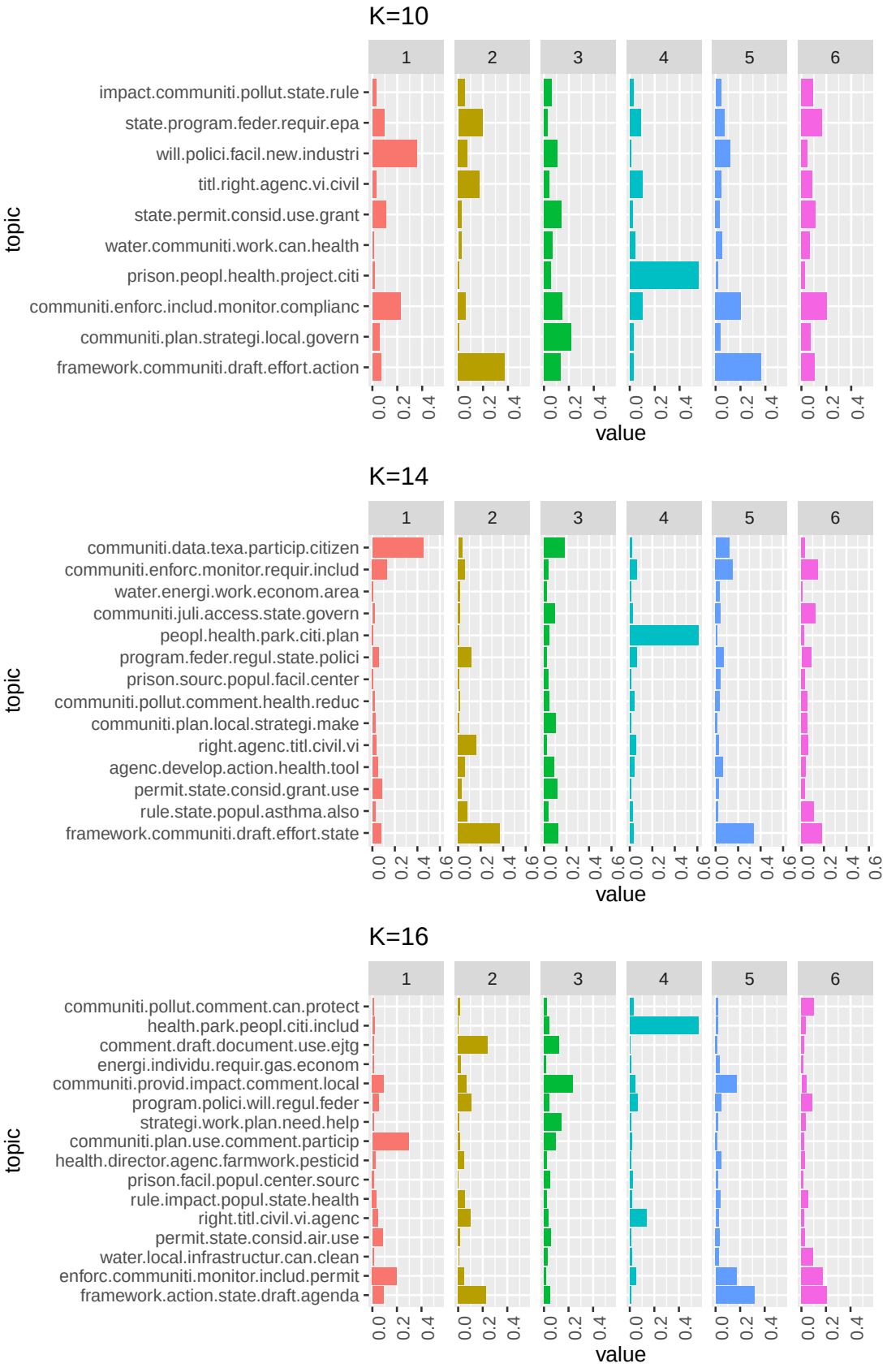
Evaluate best k value

```
docIds <- c(1, 2, 3, 4, 5, 6)
N <- length(docIds)
```

```
## -- k = 10 -- ##
# name each topic based on 1st 5 terms in topic
top5termsPerTopic_k10 <- terms(topicModel_k10, 5)
topicNames_k10 <- apply(top5termsPerTopic_k10, 2, paste, collapse = " ")

# get topic proportions from example documents
topicProportions_K10 <- theta_k10[docIds,]
colnames(topicProportions_K10) <- topicNames_k10

vizDataFrame_k10 <- melt(cbind(data.frame(topicProportions_K10),
                               document = factor(1:N)),
                         variable.name = "topic",
                         id.vars = "document")

plot_k10 <- ggplot(data = vizDataFrame_k10,
                   aes(x = topic,
                       y = value,
                       fill = document),
                   ylab = "proportion") +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = "none") +
  coord_flip() +
  facet_wrap(~ document, ncol = N) +
  labs(title = "K=10")


## -- k = 14 -- ##
# name each topic based on 1st 5 terms in topic
top5termsPerTopic_k14 <- terms(topicModel_k14, 5)
topicNames_k14 <- apply(top5termsPerTopic_k14, 2, paste, collapse = " ")

# get topic proportions from example documents
topicProportions_k14 <- theta_k14[docIds,]
colnames(topicProportions_k14) <- topicNames_k14

vizDataFrame_k14 <- melt(cbind(data.frame(topicProportions_k14),
                               document = factor(1:N)),
                         variable.name = "topic",
                         id.vars = "document")

plot_k14 <- ggplot(data = vizDataFrame_k14,
                   aes(x = topic,
                       y = value,
                       fill = document),
                   ylab = "proportion") +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
```

```
        legend.position = "none") +
  coord_flip() +
  facet_wrap(~ document, ncol = N) +
  labs(title = "K=14")


## -- k = 16 -- ##
# name each topic based on 1st 5 terms in topic
top5termsPerTopic_k16 <- terms(topicModel_k16, 5)
topicNames_k16 <- apply(top5termsPerTopic_k16, 2, paste, collapse = " ")

# get topic proportions from example documents
topicProportions_k16 <- theta_k16[docIds,]
colnames(topicProportions_k16) <- topicNames_k16

vizDataFrame_k16 <- melt(cbind(data.frame(topicProportions_k16),
                               document = factor(1:N)),
                         variable.name = "topic",
                         id.vars = "document")

plot_k16 <- ggplot(data = vizDataFrame_k16,
                   aes(x = topic,
                       y = value,
                       fill = document),
                   ylab = "proportion") +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = "none") +
  coord_flip() +
  facet_wrap(~ document, ncol = N) +
  labs(title = "K=16")
```

```
plot_k10 / plot_k14 / plot_k16
```


K=10
1
2
3
4
5
6
impact.communiti.pollut.state.rule
state.program.feder.requir.epa
will.polici.facil.new.industri
titl.right.agenc.vi.civil
state.permit.consid.use.grant
water.communiti.work.can.health
prison.peopl.health.project.citi
communiti.enforc.includ.monitor.complianc
communiti.plan.strategi.local.govern
framework.communiti.draft.effort.action
topic
0.0
0.2
0.4
value
K=14
communiti.data.texa.particip.citizen
communiti.enforc.monitor.requir.includ
water.energi.work.econom.area
communiti.juli.access.state.govern
peopl.health.park.citi.plan
program.feder.regul.state.polici
prison.sourc.popul.facil.center
communiti.pollut.comment.health.reduc
communiti.plan.local.strategi.make
right.agenc.titl.civil.vi
agenc.develop.action.health.tool
permit.state.consid.grant.use
rule.state.popul.asthma.also
framework.communiti.draft.effort.state
topic
0.0
0.2
0.4
0.6
value
K=16
communiti.pollut.comment.can.protect
health.park.peopl.citi.includ
comment.draft.document.use.ejtg
energi.individu.requir.gas.econom
communiti.provid.impact.comment.local
program.polici.will.regul.feder
strategi.work.plan.need.help
communiti.plan.use.comment.particip
health.director.agenc.farmwork.pesticid
prison.facil.popul.center.sourc
rule.impact.popul.state.health
right.titl.civil.vi.agenc
permit.state.consid.air.use
water.local.infrastructur.can.clean
enforc.communiti.monitor.includ.permit
framework.action.state.draft.agenda
topic
0.0
0.2
0.4
value

```r
svd_tsne <- function(x){
  tsne(svd(x)$u)
  }

json_k10 <- createJSON(
  phi = tmResult_k10$terms,
  theta = tmResult_k10$topics,
  doc.length = rowSums(doc_feature_matric),
  vocab = colnames(doc_feature_matric),
  term.frequency = colSums(doc_feature_matric),
  mds.method = svd_tsne,
  plot.opts = list(xlab = "", ylab = ""))

json_k14 <- createJSON(
  phi = tmResult_k14$terms,
  theta = tmResult_k14$topics,
  doc.length = rowSums(doc_feature_matric),
  vocab = colnames(doc_feature_matric),
  term.frequency = colSums(doc_feature_matric),
  mds.method = svd_tsne,
  plot.opts = list(xlab = "", ylab = ""))

json_k16 <- createJSON(
  phi = tmResult_k16$terms,
  theta = tmResult_k16$topics,
  doc.length = rowSums(doc_feature_matric),
  vocab = colnames(doc_feature_matric),
  term.frequency = colSums(doc_feature_matric),
  mds.method = svd_tsne,
  plot.opts = list(xlab = "", ylab = ""))
```

```r
serVis(json_k16)
```

Picking k values to check Based on the `FindTopicsNumber()` optimization metrics, 9, 10, 14, 16, and 20 are good candidates for k (the number of topics). Based on the plots, 10 is the highest maximum and around where the minimization plot begins to decrease more slowly. In the maximization plot, 14 is the next highest maximum peak and 16 is where the minimization plot begins to level out. We know there are 9 priority areas for the EPA and as discussed in the lab, there are 7 additional topics identified in the EPA's response to the public comments. So this would be a theoretical argument for 16 being the best k value. Based on all of the above, I chose to run three models with k equal to 10, 14, and 16.

Picking the best k value Based on the LDAvis and looking at the graphs of the top terms in each topic I would move forward with using 14 as the overall best value for k. When looking at the top term plots for k=16 as compared to k=14, there appeared to begin to be topics that were related to each other such as topics 14 and 15 in the plot that both reference community actions and participation with government. In looking at the LDAvis, the size of the circles for each topic appear to be evenly sized for k=14. Whereas, for k=16 there are more topics with smaller circles (aka less topic distribution).